Consequences of effect size heterogeneity for meta-analysis: a Monte Carlo study

Koetse, M.J.; Florax, R.J.G.M.; de Groot, H.L.F.

published in
Statistical Methods and Applications
2010

DOI (link to publisher)
10.1007/s10260-009-0125-0

document version
Publisher's PDF, also known as Version of record

Stat Methods Appl (2010) 19:217–236
DOI 10.1007/s10260-009-0125-0

ORIGINAL ARTICLE

Consequences of effect size heterogeneity for meta-analysis: a Monte Carlo study

Mark J. Koetse · Raymond J. G. M. Florax · Henri L. F. de Groot

Accepted: 18 June 2009 / Published online: 7 July 2009

Abstract In this article we use Monte Carlo analysis to assess the small sample behaviour of the OLS, the weighted least squares (WLS) and the mixed effects meta-estimators under several types of effect size heterogeneity, using the bias, the mean squared error and the size and power of the statistical tests as performance indicators. Specifically, we analyse the consequences of heterogeneity in effect size precision (heteroskedasticity) and of two types of random effect size variation, one where the variation holds for the entire sample, and one where only a subset of the sample of studies is affected. Our results show that the mixed effects estimator is to be preferred to the other two estimators in the first two situations, but that WLS outperforms OLS and mixed effects in the third situation. Our findings therefore show that, under circumstances that are quite common in practice, using the mixed effects estimator may be suboptimal and that the use of WLS is preferable.

Keywords Effect size heterogeneity · Meta-analysis · Monte Carlo analysis · OLS meta-estimator · WLS meta-estimator · Mixed effects meta-estimator · Small sample performance

JEL Classification C12 · C15 · C40

M. J. Koetse (✉) · R. J. G. M. Florax · H. L. F. de Groot
Department of Spatial Economics, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
e-mail: mkoetse@feweb.vu.nl

R. J. G. M. Florax
Department of Agricultural Economics, Purdue University, 403 W. State Street, West Lafayette, IN 47907, USA
e-mail: rflorax@purdue.edu

H. L. F. de Groot
Tinbergen Institute, Amsterdam, The Netherlands
e-mail: hgroot@feweb.vu.nl

1 Background

The stock of empirical evidence in economic research is widely scattered across journals, working papers, dissertations and unpublished manuscripts. For the majority of issues in economics, empirical studies differ largely with respect to their set-up and characteristics, and display a wide variety of outcomes, in terms of both direction and magnitude of the estimates. In the words of Hunt (1997, p. 1): "Our faith that scientists are cooperatively and steadily enlarging their understanding of the world is giving way to doubt as, time and again, new research assaults existing knowledge." The development of tools to systematically synthesise and analyse the "flood of numbers" is therefore crucial.

Next to the more or less classic narrative literature review, a method that has become increasingly popular during the last decades is meta-analysis. It is a form of quantitative research synthesis originally developed in experimental medicine, and later on extended to fields such as biomedicine and experimental behavioural sciences, specifically education and psychology. During the last two decades it has also been widely applied in many areas of economics (for recent contributions see, among many others, Abreu et al. 2005; Brons et al. 2008; De Dominicis et al. 2008; Koetse et al. 2008, 2009; Nijkamp and Poot 2004, 2005; Roberts and Stanley 2005; Weichselbaumer and Winter-Ebmer 2005). The method's intuitive appeal lies in combining often widely scattered empirical estimates and in the increase in statistical power of hypothesis testing when combining independent research results. As such, an attractive feature of meta-analysis is that pooling study outcomes provides a preferable estimate, i.e., an estimate with a smaller confidence interval. Moreover, by controlling for the differences in characteristics across studies, a meta-analysis provides a quantitative insight into which factors are relevant in explaining the variation in study outcomes. Meta-analysis thereby provides a quantitative analytical assessment of the literature, which complements the more qualitative judgment provided by a standard narrative literature review (Stanley 2001).

Although there has been a wide increase in the application of meta-regression analysis in economics, the method still faces various methodical difficulties. For example, in economics, data constraints as well as the desire to be 'different' lead to varying sets of control variables across studies, inducing omitted variable bias in at least a subset of the existing empirical studies. Moreover, since the true data generating process is generally unknown, different effect size measures are reported in primary studies and these are pooled in a meta-analysis sample.

In Koetse et al. (2005) Monte Carlo techniques are used in order to investigate the consequences of these two problems. The results of this study show that misspecifications in primary studies carry over to meta-analysis results. Specifically, the results show that a meta-estimator that does not account for these issues is not useful since it is biased and virtually always rejects the null hypothesis that there is no real effect, regardless whether this is true or not. However, the current practice of accounting for these issues by means of dummy variables goes a long way in mitigating their negative effects. Of course, in actual practice it may be the case that some sources of omitted variable bias cannot be identified. However, since primary model specifications are clearly observed and any omissions compared to other model specifications can be

controlled for, this will not occur very often in our view. Therefore, any systematic effects in underlying studies caused by misspecification, or by any other differences in study characteristics, can generally be picked up by a meta-analysis. Of course, when all underlying primary models suffer from omitted variables, a meta-analysis clearly cannot take this into account. It may also be unclear which model specification in primary studies is preferable, i.e., it may be difficult to judge whether differences in model specification actually represent omitted variables or not. However, these potential problems affect the entire literature, not only the results of a meta-analysis, implying they are of a more general and fundamental nature.

In the current study we provide a more general analysis and aim to analyse the impact of effect size heterogeneity on the results of a meta-analysis. Specifically, we investigate heterogeneity in effect size precision (heteroskedasticity) and two types of random effect size variation. The difference between these two types is subtle and will be discussed in detail in the next section. We address the small sample behaviour of three meta-regression estimators, i.e., the OLS estimator, the weighted least squares (WLS) estimator and the mixed effects estimator, in the presence of the three types of effect size heterogeneity. We use the bias and mean squared error (MSE) of the estimators and the size and power of the statistical tests as performance indicators.

The remainder of this article is organised as follows. The next section discusses three sources of effect size heterogeneity in more detail. Section 3 describes the experimental design, while in Sect. 4 we present and discuss the simulation results. In Sect. 5 we analyse the effects of increasing the sample size of both primary studies and meta-analysis in order to draw inferences on the asymptotic properties of the meta-estimators. Section 6 concludes.

2 Sources and characteristics of effect size heterogeneity

Heterogeneity in effect size precision and random variation in the true underlying effect across primary studies may have substantial consequences for the results of a meta-analysis. To illustrate the potential problems, let T_s be the estimate of the true effect size θ_s from primary study s. This estimate is generally assumed to be normally distributed, such that:

$$T_s \sim N\left(\theta_s, \sigma_s^2\right), \tag{1}$$

where σ_s^2 represents the estimates' variance. Estimate variance generally displays large heterogeneity, causing heteroskedasticity in a meta-analysis sample. Important sources of heteroskedasticity are differences in primary study sample size and differences in model specification and data type. Ultimately, the consequences for meta-analysis are potentially serious. Crucial is the fact that, assuming a standard OLS estimation, effect sizes with a higher variance get as much weight as effect sizes with a lower variance. Therefore, OLS is not efficient, i.e., does not attain the minimum estimated variance, and the variance estimator is biased (see Stanley and Jarrell 1989). The optimal way to correct for this problem is to weight the effect sizes with their respective standard errors. Since the actual standard errors are unknown in practice, meta-analysis

commonly uses the standard errors estimated by the primary model, which is a good approximation unless sample sizes in primary studies are exceptionally small (see Hedges 1994, p. 287).

A second issue is related to the characteristics of the true underlying effect size θ_s. After the *systematic* variation in effect sizes is controlled for by including dummy variables in the meta-model specification, basically two assumptions on the nature of the remaining *non-systematic* effect size variation exist. An often used assumption is that effect size variation is due solely to sampling error and that the true effect size θ_s is constant across primary studies, i.e., $\theta_s = \theta$. An alternative assumption is that the remaining variation is partly due to random variation in the true underlying effect across primary studies, such that:

$$\theta_s \sim N\left(\theta\ , \tau^2\right), \tag{2}$$

where τ^2 represents the variance of the population effect size, generally referred to as the between-study variance. The standard assumption in meta-analysis is that possible random variation holds for all effect sizes. However, a probably more plausible assumption is that differences between primary studies, such as differences in data type, econometric technique and model specification, cause random variation in only a subset of the meta-analysis sample. For instance, it is very likely that the bias in effect sizes due to omitted variables in primary studies is different for every primary study. This means that part of the impact of omitted variables is systematic and may be picked up by a dummy variable, and part of the impact is random. The difference between random variation due to omitted variables and random variation in the true underlying effect is not related to the source of random variation. In fact, after controlling for the systematic part of the effect size variation, the result in both situations is a random effect size distribution around zero. The difference lies in the fact that random variation in the true underlying effect causes randomness of each effect size in the meta-analysis sample, whereas random variation due to omitted variables (or due to differences in data type, functional form, level of data aggregation, etc.), only causes randomness of effect sizes from misspecified primary studies. This may have serious consequences for the optimal weight structure of a meta-estimator, implying that the two sources of random effect size variation may have different consequences for the results of a meta-analysis. Since these issues are difficult to address analytically, we use Monte Carlo simulations for analysing the impact of effect-size heterogeneity on meta-analysis and meta-estimator performance.

3 Experimental design

In this section we discuss the experimental design that serves as the basis for our simulations. The data generation process (DGP) consists of four steps: (1) generating the primary data; (2) estimating the primary models; (3) performing the meta-analyses using the estimated effect sizes and characteristics of the primary studies as inputs;

(4) analysing the small sample performance of the meta-estimators. These four steps are discussed in detail below.[1]

3.1 Generating the primary data

As true underlying primary model we use an unrestricted Cobb–Douglas function of the form:

$$y = e^{\alpha} x^{\beta_0} z^{\beta_1} e^{\varepsilon}, \tag{3}$$

where y is a stochastic dependent variable, x and z are exogenous variables, α, β_0 and β_1 are parameters, and ε is an error term. Without loss of generality we set both α and β_1 equal to 1, while the error term ε is normally distributed with mean 0 and variance σ^2. This error term is regenerated for each replication. In our model, β_0 is the parameter of interest, i.e., the true underlying effect. We draw β_0 randomly from a normal distribution with mean μ and between-study variance τ^2, and set μ equal to 1 and 0 in order to analyse the cases with and without a true underlying relationship. Furthermore, if not mentioned otherwise, the sample size of the primary model is fixed at 500 and the number of replications for each primary study combination is 5,000. The variable x is generated, once, according to:

$$x = e^{\vartheta}, \tag{4}$$

where ϑ is drawn from a uniform (0,1) distribution. In order to be able to induce omitted variable bias in a primary study (see Sect. 3.2) we relate x to z by generating z according to:

$$z = x^{\lambda} e^{\psi}, \tag{5}$$

where λ is a parameter and ψ is an error term drawn, once, from a uniform (0,1) distribution (ψ, ϑ and ε are independent). Note that the potential bias induced in the estimate of β_0 when z is excluded from the primary model does not only increase with the correlation coefficient between x and z, but also with the variance of z (see Koetse et al. 2005). Obviously, when $\lambda = 0$, the correlation between x and z is zero, implying that the bias in β_0 when z is excluded from the primary model is zero as well. However, when we increase the value of λ, both the correlation between x and z and the variance of z are increased, thereby invariantly increasing the bias in the estimate of β_0.

The main issues analysed in this article revolve around effect size heterogeneity. First, we increase the degree of heteroskedasticity via the error term in primary studies. This error term is normally distributed with mean 0 and variance σ^2, which we vary systematically between 1 and 10 with increments of 1. Second, we investigate the consequences of random variation in the true underlying effect, implying that $\tau^2 > 0$.

[1] The computer programs used for the analyses in this article are written in Gauss 8.0.

Note that τ^2 is fixed within a meta-analysis, implying not that β_0 is fixed, but that the distribution from which the true underlying effect is drawn is identical for each effect size within a single meta-analysis. Ultimately, we vary τ^2 systematically across (not within) meta-analyses, varying its value between 0 and 2 with increments of 0.2. Finally, we abandon the standard assumption in meta-analysis that random effect size variation holds (equally) for each effect size in a meta-analysis sample. Instead we assume that differences between primary studies induce random variation in only a subset of the sample. We replicate this situation by creating a non-systematic impact of omitted variables across primary studies by systematically varying λ, the parameter that determines the amount of bias due to omitted variables in a primary study. Specifically, we draw λ from a normal distribution with mean 1 and variance υ^2, which is larger than zero when part of the impact of omitted variables is random. Further details on the experimental design of the simulations are given in the relevant subsections of Sect. 4.

3.2 Estimating the primary models

Our approach is different from most Monte-Carlo studies in meta-analysis (see, e.g., Bijmolt and Pieters 2001; Field 2001; Kuhnert and Böhning 2007; Oswald and Johnson 1998; Sanchez-Meca and Marin-Martinez 1997, 1998) in that we explicitly incorporate the stage of the primary data analysis (see Stanley 2008, for a similar approach). Besides the fact that this allows us to introduce omitted variable bias in primary studies, we may also introduce erroneous effect size operationalisations and assess their impact on the results of a meta-analysis. Specifically, we use the data generated by the model in Eq. (3) to estimate a log-linear model, which is mathematically equivalent to the model in (3), and an alternative linear model.[2] The log-linear model is given by:

$$\ln(y) = \alpha + \beta_0 \ln(x) + \beta_1 \ln(z) + \varepsilon. \tag{6}$$

We estimate this model by OLS, which produces $\hat{\alpha}$, $\hat{\beta}_0$ and $\hat{\beta}_1$ as estimates of α, β_0 and β_1, respectively. The parameter of interest is the double-log elasticity of $\ln(y)$ on $\ln(x)$, given by $\eta = \hat{\beta}_0$. This elasticity is correctly estimated given our data generating process. The standard error of the elasticity is simply the standard error of $\hat{\beta}_0$. In order to induce omitted variable bias we use two primary model specifications, i.e., the correctly specified primary model in Eq. (6) and a misspecified version of this model from which $\ln(z)$ is excluded as an explanatory variable. The latter model induces omitted variable bias in $\hat{\beta}_0$ when $\lambda \neq 0$ (see Koetse et al. 2005).

An alternative elasticity estimate is obtained by estimating the linear primary model specification:

$$y = \alpha^* + \beta_0^* x + \beta_1^* z + \varepsilon^*. \tag{7}$$

[2] Of course, the choice of the true underlying model is arbitrary, i.e., we could also have chosen the linear model as the true underlying model. However, there is no reason to suspect that the results presented later on in this article would change fundamentally when our choice of true underlying model would have been different.

Using OLS to estimate this model produces $\hat{\alpha}^*$, $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ as estimates of α^*, β_0^* and β_1^*, respectively. In this linear model we estimate the intrinsically non-linear relationship between y, x and z, and compute a point-elasticity at the sample mean, for say primary study m, as $\eta_m = \hat{\beta}_{0m}^*(\bar{x}_m/\bar{y}_m)$. In reality the estimation of different effect size measures may occur frequently, simply because the true underlying model is unknown and researchers may build their analysis on an erroneous model specification. The argument for using the ratio of mean values as the evaluation point is that this is common practice in most studies.[3] To calculate the standard error of this elasticity we use the Delta method (see Greene 2000, pp. 359–360). As before, in order to induce omitted variables bias we use the model specification in Eq. (7) and a specification from which z is excluded as an explanatory variable. The latter model induces omitted variable bias in $\hat{\beta}_0^*$ when $\lambda \neq 0$.

3.3 Specification of the meta-estimators

The primary aim of this study is to compare the small sample performance of three well-known meta-estimators under the three regimes of effect size heterogeneity introduced in Sect. 3.1. The effect sizes in our meta-analyses are the elasticities produced by the primary model estimations. The amount of primary study misspecification in a meta-analysis sample is set at a moderate level; both the proportion of point-elasticities and the proportion of effect sizes from studies with omitted variables bias in the meta-analysis is fixed at 50%. We furthermore perform separate analyses for $\mu = 0$ and $\mu = 1$. Within these restrictions the elasticities are randomly sampled from the 5,000 primary study replications. Finally, the meta-analysis sample size is 50 and the number of meta-analysis replications is equal to 5,000. Our first model is a meta-regression model with dummy variables in order to correct for primary study misspecifications. This model is given by:

$$\eta_s = \delta_0^{ols} + \delta_1^{ols} D_s^{pe} + \delta_2^{ols} D_s^{ov} + \xi_s^{ols}, \tag{8}$$

where η_s is a vector of elasticities, D_s^{pe} is a dummy variable equal to one if the elasticity is a point-elasticity, D_s^{ov} is a dummy variable equal to one if the primary study is estimated without z among the explanatory variables, and ξ_s^{ols} is an error term. The model is estimated by OLS, producing $\hat{\delta}_0^{ols}$ as an estimate of the true underlying effect μ, and $\hat{\delta}_1^{ols}$ and $\hat{\delta}_2^{ols}$ as estimates of the dummy variables that should pick up the systematic impact of point-elasticities and omitted variable bias.[4]

We subsequently test the performance of the WLS meta-regression estimator proposed by Stanley and Jarrell (1989). This estimator accounts for inherent

[3] A common alternative is to use the median of x and y.

[4] Although most meta-analyses in economics apply the weighted least squares model in (9), the OLS model in (8) may be used when the standard errors of study outcomes are not available (e.g., in the contingent valuation literature). Another procedure that is often used in this situation is to weight the data with the square root of primary study sample size. This is suboptimal but at least the sample size is strongly related to the standard error. The results obtained for the OLS estimator therefore provide information about the lower bound performance for this type of studies.

heteroskedasticity in meta-analysis by weighting the meta-analysis data with a measure of effect size precision, the ideal measure being the square root of the effect size variance. However, since these are unknown, the estimated effect size standard errors are generally used for this purpose. The WLS estimator reads as:

$$\eta_s/w_s = \delta_0^{wls}\,(1/w_s) + \delta_1^{wls}\left(D_s^{pe}/w_s\right) + \delta_2^{wls}\left(D_s^{ov}/w_s\right) + \xi_s^{wls}/w_s, \qquad (9)$$

where w_s is the weight of the effect size from study s, i.e., the standard error of the elasticity, and ξ_s^{wls} is an error term.[5] The model in (9) is estimated by OLS, producing $\hat{\delta}_0^{wls}$ as an estimate of the true underlying effect μ, and $\hat{\delta}_1^{wls}$ and $\hat{\delta}_2^{wls}$ as estimates of the dummy variables.[6]

The third estimator is the mixed effects estimator. The difference between the mixed effects and WLS estimators is that the latter assumes that the true underlying effect is fixed, whereas the former allows for random variation of the true effect across primary studies, and assumes that it is drawn from a population of effect sizes with mean μ and between-study variance τ^2. The mixed effects model makes an explicit distinction between effect size variance and between-study variance, which has obvious consequences for the model's weight structure. Since the between-study variance τ^2 is unknown it has to be estimated by the model. For this purpose we use a maximum likelihood estimator (see Sutton et al. 2000; Brockwell and Gordon 2001). The log-likelihood is given by:

$$\text{LogL} = -0.5\sum_{s=1}^{S}\left[\left(\eta_s - \delta_0^{me} - \delta_1^{me}D_s^{pe} - \delta_2^{me}D_s^{ov}\right)^2/\left(\tau^2 + w_s^2\right) + \ln\left(\tau^2 + w_s^2\right)\right]. \qquad (10)$$

The coefficient of interest is $\hat{\delta}_0^{me}$, which is an estimate of the mean μ of the underlying population effect size, while $\hat{\delta}_1^{me}$ and $\hat{\delta}_2^{me}$ are estimates of the dummy variables and $\hat{\tau}^2$ is an estimate of the population effect size variance τ^2. Observe that the mixed effects estimator reduces to the WLS estimator when $\hat{\tau}^2 = 0$.

[5] There is some confusion in the literature on whether to use the standard errors or the variances as weights. The fixed effects model, in which a weighted mean effect size is calculated, uses the variances as weights (see Hedges 1994, pp. 287–288), while in a regression context the standard errors are used (see Stanley and Jarrell 1989). These procedures give identical results, because OLS minimizes the squared errors, thereby squaring the standard errors. Calculating a weighted mean with variances as weights (fixed effects model), and using WLS estimation with only a constant using standard errors as weights, produces identical estimates. The only difference between fixed effects and WLS is that their variance estimators are different; see also footnote 6.

[6] A slightly different model is the fixed effects regression estimator, which is different from the model in (9) in that it assumes that study outcomes display no excess variation. For this purpose a modification of the standard errors from the model in (9) is necessary. The fixed effects standard errors are given by $stderr/\sqrt{msr}$, where *stderr* is the standard error of the meta-analysis estimate given by the computer program and *msr* is the mean squared residual of the meta-analysis (see Hedges 1994, p. 296). For the parameter values in our simulation exercises this implies that the WLS model in (9) produces more conservative standard errors than the fixed effects regression model, i.e., WLS produces wider confidence intervals. Otherwise the models are identical.

3.4 Assessing small sample performance

The central issue in this study is how well the meta-estimators recover the value of the population effect size μ, in terms of both size and statistical significance. The parameters of interest are therefore the true underlying effect μ and the meta-estimates $\hat{\delta}_0^{ols}$, $\hat{\delta}_0^{wls}$ and $\hat{\delta}_0^{me}$. Effect size heterogeneity may affect the estimators on several dimensions, so we use three different performance indicators to investigate its impact. First, the bias (BIAS) of the estimators measures the difference between the average value of the estimates and μ. Although the impact of effect size heterogeneity may average out, in which case estimator bias is equal to zero, the variance of the estimators may still be affected. We therefore also use the MSE of the estimate as a performance indicator. This second indicator combines the bias and the variance of the estimators, and measures the average distance of the estimate to the true parameter, i.e., the smaller the MSE, the closer the estimate will be to the true parameter, on average. The third and final indicator is the proportion of statistically significant results (SIG) of the meta-estimators. Formally, for $\hat{\delta}_0^{ols}$ these indicators are given by:

$$\text{BIAS}\left(\hat{\delta}_0^{ols}\right) = \text{E}\left(\hat{\delta}_0^{ols} - \mu\right) \approx \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\delta}_0^{ols} - \mu\right)_r, \tag{11}$$

$$\text{MSE}\left(\hat{\delta}_0^{ols}\right) = \text{E}\left(\hat{\delta}_0^{ols} - \mu\right)^2 = \text{BIAS}\left(\hat{\delta}_0^{ols}\right)^2 + var\left(\hat{\delta}_0^{ols}\right) \approx \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\delta}_0^{ols} - \mu\right)_r^2, \tag{12}$$

$$\text{SIG}\left(\hat{\delta}_0^{ols}\right) = \frac{1}{R}\sum_{r=1}^{R} I\left(|t_{n-k}| > t_{crit}\right)_r, \tag{13}$$

where $r = 1, 2, \ldots, R$ indexes the meta-analyses replications.[7] In Eq. (13) I is an indicator function equal to one if the absolute t value of the meta-estimate is greater than a pre-specified critical t value, denoted by t_{crit}, and 0 otherwise. We apply two-sided significance tests using a 5% significance level. When $\mu = 0$ and $H_0: \mu = 0$, we are interested in the probability of a Type I error, i.e., the probability that the statistical test on the meta-estimate erroneously rejects H_0. Therefore, when $\mu = 0$, SIG corresponds to the proportion of Type I errors. From now on we will refer to this as the size of the statistical test on the meta-estimates. Alternatively, when $\mu = 1$, and under the same null-hypothesis, we are interested in the probability of a Type II error, i.e., the probability that the statistical test on the meta-estimate erroneously accepts H_0. When $\mu = 1$, SIG corresponds to (1—probability of a Type II error), or the power of the statistical test. Since erroneously rejecting the null-hypothesis requires a considerably larger confidence interval than erroneously accepting the null-hypothesis, the two indicators provide different types of information on statistical significance. This is the most important reason why we distinguish between a zero ($\mu = 0$) and a non-zero ($\mu = 1$) true underlying effect. However, the two tests are clearly related, since

[7] The performance indicators for $\hat{\delta}_0^{wls}$ and $\hat{\delta}_0^{me}$ are obtained by replacing $\hat{\delta}_0^{ols}$ by $\hat{\delta}_0^{wls}$ and $\hat{\delta}_0^{me}$ in Eqs. (11)–(13).

decreasing standard errors simultaneously cause a decrease in size and an increase in power, *ceteris paribus*.

4 Simulation results

In this section we analyse the performance of the three meta-estimators under various forms of effect size heterogeneity. In Sect. 4.1 we analyse the impact of increasing degrees of heteroskedasticity in the meta-analysis sample. We address the consequences of increasing between-study variance in Sect. 4.2, while Sect. 4.3 investigates the impact of non-systematic effects of omitted variables in primary studies. The latter basically induces a form of between-study variance, the difference being that the random variation does not apply to the entire sample but only to those effect sizes that are obtained from misspecified studies. Under these circumstances all three estimators use erroneous weighting structures, implying that it is unclear a priori which estimator is to be preferred.

4.1 Increasing degrees of heteroskedasticity

In this section we analyse the impact of increasing primary study error variance and of increasing heteroskedasticity on the results of a meta-analysis. In the experimental design we only vary the primary study error variance and keep constant all other parameters. Specifically, the between-study variance $\tau^2 = 0$ and omitted variable bias is constant across primary studies, i.e., $\lambda = 1$ and $\upsilon^2 = 0$, in which case the mixed effects estimator should reduce to the WLS estimator. Primary studies are estimated with an error variance ranging from 1 to 10, with increments of 1. For these error variance values, correctly specified primary studies display average R^2 values ranging from 0.15 to 0.02 when $\mu = 0$, and from 0.30 to 0.04 when $\mu = 1$. These R^2 values are reasonable compared to the values found in many areas of economic research.

In the graphs below we show the bias, MSE or size/power along the vertical axis. Along the horizontal axis we measure the degree of heteroskedasticity. We distinguish between ten cases. The first case is the case with no heteroskedasticity; all effect sizes in the meta-analysis are drawn from primary studies with error variance 1. From the second case up to the tenth case we systematically increase the average error variance and the degree of heteroskedasticity, by systematically increasing the proportion of effect sizes drawn from studies with a higher error variance by 10%. In Table 1 we present the resulting proportions of effect sizes drawn from studies with a pre-specified error variance for each of the ten cases. Note that for each case both the average effect size variance and the degree of heteroskedasticity are higher than in the previous cases.

In Fig. 1 we present the performance of the three estimators on the three indicators for these ten cases. As expected, the WLS and the mixed effects estimators produce almost identical results, and the bias of the three estimators is not affected by heteroskedasticity and increasing error variance. Estimator variance clearly increases, however. Most interesting is that with increasing severity of the heteroskedasticity the variance of OLS deteriorates rapidly *vis-à-vis* the variance of WLS and the mixed

Table 1 Proportion of effect sizes from primary studies with a pre-specified error variance in ten different cases (in %)

Case	Value of error variance									
	1	2	3	4	5	6	7	8	9	10
1	100	–	–	–	–	–	–	–	–	–
2	90	10	–	–	–	–	–	–	–	–
3	80	10	10	–	–	–	–	–	–	–
4	70	10	10	10	–	–	–	–	–	–
5	60	10	10	10	10	–	–	–	–	–
6	50	10	10	10	10	10	–	–	–	–
7	40	10	10	10	10	10	10	–	–	–
8	30	10	10	10	10	10	10	10	–	–
9	20	10	10	10	10	10	10	10	10	–
10	10	10	10	10	10	10	10	10	10	10

effects estimator. However, judging by the size, this is more than compensated by the fact that OLS produces wider confidence intervals, a fact that appears to leave the power unaffected.[8] In conclusion, since OLS is highly inefficient in the presence of heteroskedasticity, not having the standard errors of effect sizes in a meta-analysis leaves OLS unbiased but substantially increases its variance, and as such may cause statistical inferences to be substantially off the mark.

4.2 Random variation of the true underlying effect

In this section we introduce a random effect size and systematically increase the variance of the population effect size. Specifically, we increase the between-study variance τ^2 from 0 to 2 with increments of 0.2. With respect to heteroskedasticity we replicate the situation of the tenth case in the previous section, i.e., maximum heteroskedasticity. Values of other variables and parameters remain unchanged. The results are presented in Fig. 2.

In general, and according to expectations, increasing between-study variance has no impact on estimator bias but it increases estimator variance substantially, as demonstrated by the increase in MSE of all three estimators. WLS uses erroneous weights when the between-study variance is larger than zero, and although the estimated variance is biased downwards in this case, the effects on the WLS estimator variance are not clear a priori. Our results clearly show that WLS estimator variance increases *vis-à-vis* the variance of the OLS and mixed effects estimators. Together with the narrow confidence intervals produced by this estimator this also causes an increase in size. The narrow confidence intervals also cause the WLS power to be somewhat higher than the power associated with the OLS and mixed effects

[8] See also Higgins and Thompson (2004) for an analysis of Type I error rates on *non-relevant* study characteristics under various sources of effect size heterogeneity.

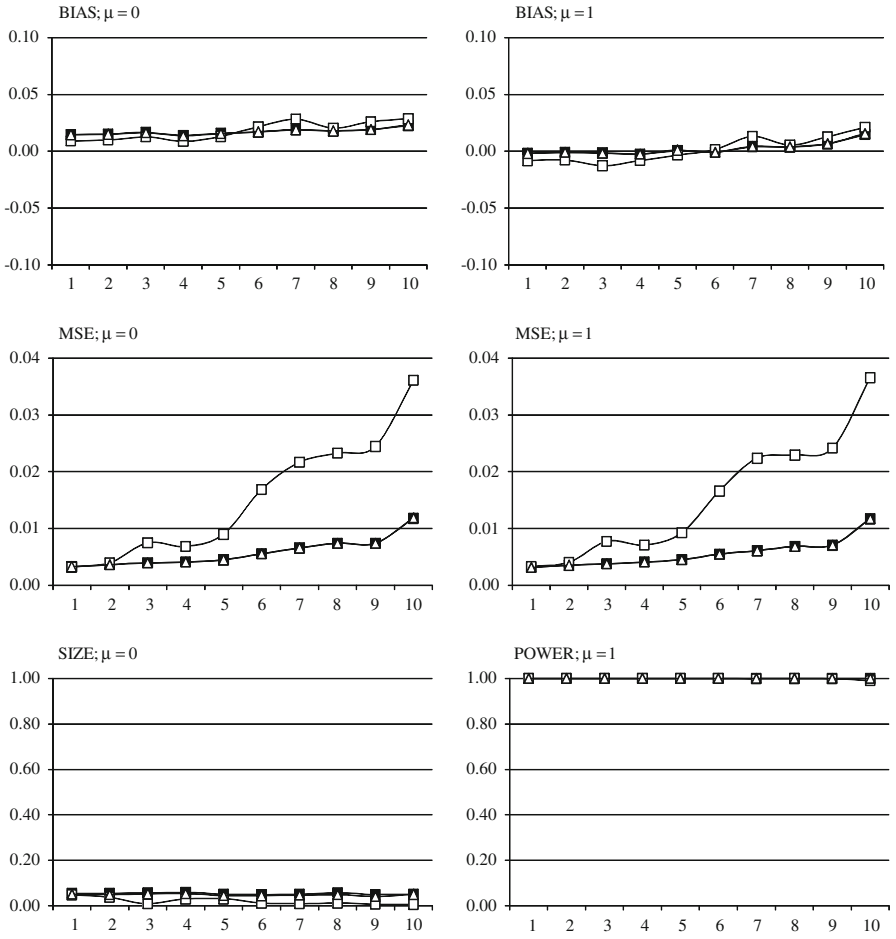

Fig. 1 Impact of heteroskedasticity on meta-estimator performance. Along the vertical axis are the BIAS (*top*), MSE (*middle*), and SIG (*bottom*) for the case where the fixed population effect size $\mu = 0$ (*left*) and $\mu = 1$ (*right*), against the degree of heteroskedasticity in the meta-sample along the horizontal axis. The different lines pertain to the OLS (*open square*), the WLS (*filled square*) and the mixed effects (*open triangle*) meta-estimators. See main text for further details

estimators, despite the fact that the WLS variance is higher than the OLS and mixed effects counterparts.[9]

A somewhat surprising result at first sight is that the MSE of and the size associated with the OLS estimator slowly converge to their mixed effects counterparts. Although this may seem strange, the result follows directly from a comparison of the weight structures used in the estimators. When between-study variance increases, its

[9] For robustness we also tested the situation with no heteroskedasticity; the patterns are similar but the differences are smaller. The bias remains unaffected, but the variance of the WLS estimator is higher than for the OLS and mixed effects estimators. The inflation in size associated with WLS is no longer absent, however.

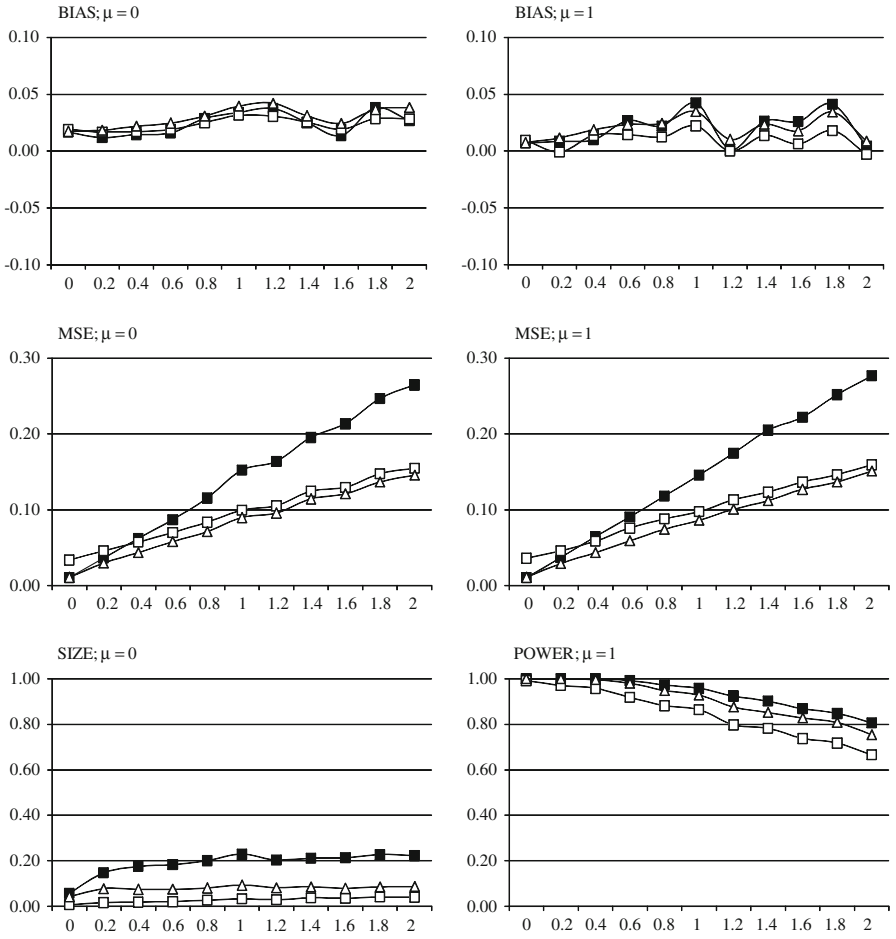

Fig. 2 Impact of random variation in the true underlying effect on meta-estimator performance. Along the vertical axis are the BIAS (*top*), MSE (*middle*), and SIG (*bottom*) for the case where the mean random effect size $\mu = 0$ (*left*) and $\mu = 1$ (*right*), against between-study variance τ^2 along the horizontal axis in absolute values. The different lines pertain to the OLS (*open square*), the WLS (*filled square*) and the mixed effects (*open triangle*) meta-estimators. See main text for further details

magnitude relative to effect size variance increases as well. As a consequence, effect size variance becomes less and less important in the weight structure of the mixed effects model. The central point is now that between-study variance is equal for each effect size in the meta-analysis, implying that, under increasing between-study variance, the weight structure of the mixed effects model tends towards a structure in which each effect size gets an equal weight. Since the OLS estimator gives each effect size an equal weight by definition, the estimates produced by the two estimators converge under increasing between-study variance. Also the size associated with OLS is smaller than its mixed effects counterpart. Since the MSE of the OLS estimator is higher in all circumstances, this implies that OLS confidence intervals are substantially wider than mixed effects confidence intervals.

Clearly, when the true underlying effect is randomly distributed across primary studies, the mixed effects estimator is to be preferred to the OLS and also to the WLS estimator. Although the bias of all three estimators remains unaffected, the WLS variance increases sharply *vis-à-vis* the variance of the mixed effects but also the OLS estimator. Moreover, the size associated with WLS increases for increasing values of the variance of the population effect size.

4.3 Random variation due to omitted variables

The question is now whether the insights obtained in the previous section also hold when slightly different but more plausible assumptions are made regarding random effect size variation. As discussed in Sect. 2, random effect size variation may be caused by other factors than pure random variation of the true effect across all primary studies. For instance, it may be that differences between underlying studies with respect to data type or econometric technique cause random variation of the true underlying effect for only part of the effect sizes included in a meta-analysis. Under these circumstances all three estimators use erroneous weights, implying that it is unclear a priori which estimator is to be preferred.

To analyse the consequences of this situation we change the assumption that the effect of omitted variables is constant across primary studies. The necessary conditions for this assumption to hold in reality are implausible at best. For each primary-study replication we draw λ, the parameter that determines the amount of bias due to omitted variables in primary studies, from a normal distribution with mean 1 and variance υ^2, which we fix at 4. This means that part of the omitted variable bias is systematic, which should be picked up by the dummy variable D^{ov}, and that part of the bias in the meta-analysis sample is random. The difference between random effect size heterogeneity due to omitted variable bias and the random effect size heterogeneity introduced in the previous section is not due to the fact that the sources of random effect size variation are different. In fact, after controlling for the systematic part of the effect size variation, the result in both situations is a random effect size distribution around zero (see also Stanley 2008).[10] The difference lies in the fact that random variation in the true underlying effect causes randomness of each effect size in the meta-analysis sample, whereas random variation due to omitted variable bias only causes randomness of effect sizes from misspecified primary studies. Potential differences between the two sources of random effect size variation should therefore show up when we vary the proportion of effect sizes with omitted variable bias in the meta-analysis sample.[11]

[10] The systematic part of the variation under random effect size heterogeneity due to omitted variable bias is picked up by D^{ov}, while under random variation of the true underlying effect it is picked up by the constant in the meta-model.

[11] Note that the results and patterns identified in the previous section do not change when we vary the proportion of effect sizes with omitted variable bias. Therefore, if the patterns found in this section are dependent on this proportion, we can conclude that the two sources of random effect size variation have different consequences for the small sample performance of the three meta-estimators.

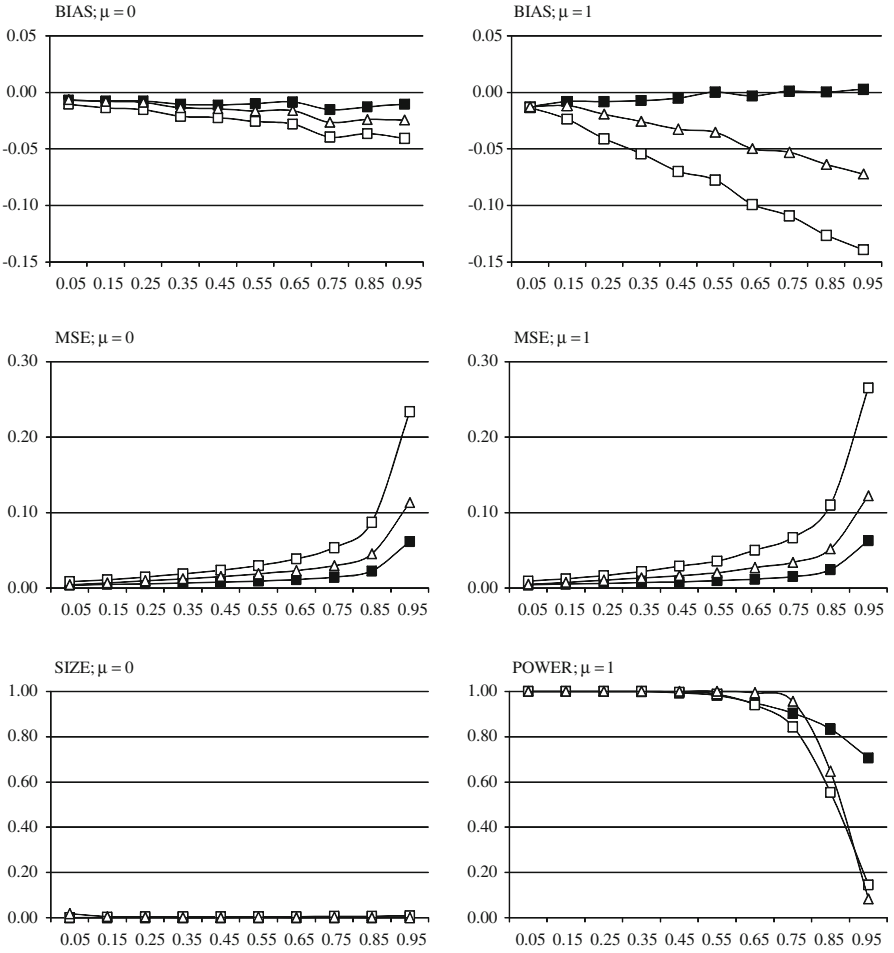

Fig. 3 Impact of random effect-size variation due to omitted variables on meta-estimator performance. Along the vertical axis are the BIAS (*top*), MSE (*middle*), and SIG (*bottom*) for the case where the fixed population effect size $\mu = 0$ (*left*) and $\mu = 1$ (*right*), against an increasing proportion in the meta-analysis sample of effect sizes from primary model specifications with omitted variables. The different lines pertain to the OLS (*open square*), the WLS (*filled square*) and the mixed effects (*open triangle*) meta-estimators. See main text for further details

We systematically increase this proportion from 0.05 to 0.95 with increments of 0.1. Heteroskedasticity is absent and we set the error variance at 5, between-study variance $\tau^2 = 0$, and the meta-analysis sample size is now equal to 150 for practical purposes. Values of other variables and parameters remain unchanged. Results are shown in Fig. 3.

The bias of the WLS estimator is clearly unaffected and lower than that of the OLS and mixed effects estimators; the bias of the latter two increases under increasing proportions of effect sizes from misspecified studies, especially when the true underlying

effect is equal to one.[12] Also, the OLS variance is substantially higher than the variance of the other two estimators. Another interesting insight, and in clear contrast to the results reported in the previous section, is that the variance of the mixed effects estimator is higher than its WLS counterpart. Under this regime of effect size heterogeneity, the mixed effects estimator erroneously assigns the estimated between-study variance to all estimates, and effect sizes from correctly specified primary models get a weight that is too low. Finally, the size of the statistical tests for all three estimators is small, while the power decreases rapidly when high proportions of effect sizes from misspecified underlying studies are included, especially for the OLS and mixed effects estimators.[13]

Given the fact that the source of random effect size variation is not known in empirical applications, our findings show that, under circumstances that are not uncommon in reality, using the mixed effects estimator is likely suboptimal, and using WLS is clearly preferable. It would be interesting to test the performance of an estimator that allows for a random effect but only for those effect sizes from misspecified primary models. However, in practice such an estimator would be of little value since the source of random variation is unknown. Moreover, the possible sources of random variation would generally be many (data type, econometric technique, omitted variables, etc.), implying that the meta-estimator would become increasingly complex and that, in most applications, the model would no longer be identified or it would no longer converge.

5 Differential effects of primary study and meta-analysis sample size

Despite the fact that various meta-estimators have been developed to control for the potential negative consequences of effect size heterogeneity, it is clear from the previous section that it still has negative effects on meta-estimator performance. However, one of the appealing features of meta-analysis is its integrative nature, implying strong improvement in results as sample size increases. Since both primary study sample size and meta-analysis sample size may go to infinity, there are two types of asymptotics to meta-estimators (see Hedges and Olkin 1985). Although the total sample size, i.e., the sum of all primary study sample sizes, may remain unchanged, primary study and meta-analysis sample size may have totally different effects on the results of a meta-analysis. Since our simulation design allows us to vary both sample sizes it is very well suited to analyse this specific issue.

[12] This is a rather surprising result since the variation from omitted variables is random, not systematic. In fact, when we reduce the proportion of point-elasticities in the meta-analysis, this pattern disappears. The patterns with respect to the MSE and the size and power of the tests remain unchanged. Clearly, therefore, we are picking up an interaction effect of point-elasticities and omitted variables where estimator bias is concerned, implying this particular result is not generally applicable in the situation analysed in this section. Still, since the inclusion of erroneous effect sizes measures in meta-analysis is common, WLS may be less biased than OLS and mixed effects in many applications in economics.

[13] For robustness we also kept the proportion of effect sizes from misspecified models constant at 50% and systematically increased the value of λ from 0 to 4 with increments of 0.4. The patterns are almost identical to the patterns observed in Fig. 3. Results are available upon request from the authors.

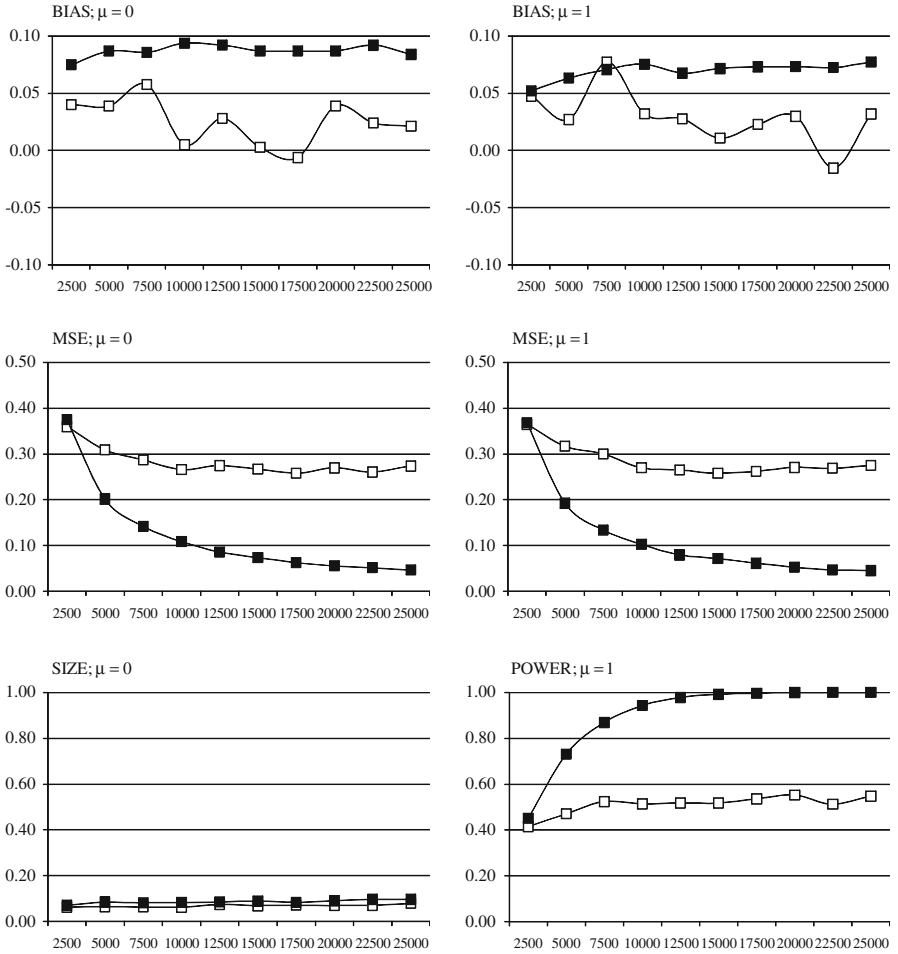

Fig. 4 Impact of sample size on meta-estimator performance. Along the vertical axis are the BIAS (*top*), MSE (*middle*), and SIG (*bottom*) for the case where the fixed population effect size $\mu = 0$ (*left*) and $\mu = 1$ (*right*), against the total number of underlying observations. The different lines pertain to the WLS meta-estimator under increasing primary study sample size (*open square*) and increasing meta-analysis sample size (*filled square*). See main text for further details

First, we systematically increase the sample size of the primary studies from 100 to 1000, with increments of 100, and fix the meta-analysis sample size at 25. Second, we increase meta-analysis sample size systematically from 25 to 250, with increments of 25, and keep primary study sample size fixed at 100. We can now distinguish between ten cases with varying primary study and meta-analysis sample size, but with an equal number of total underlying observations in each case. Specifically, with each case the number of total underlying observations increases with 2,500 under both regimes. This way we can clearly observe the differential impact of meta-analysis sample size and primary study sample size on the results of a meta-analysis. For simplicity we only present results for the WLS estimator, since the patterns for the other two

meta-estimators are identical.[14] Heteroskedasticity is absent and error variance and between-study variance are equal to 5 and 2, respectively. We keep the impact of omitted variables fixed across primary studies ($\lambda = 1$ and $\upsilon^2 = 0$) and both the proportion of effect sizes from misspecified studies and the proportion of point-elasticities in the meta-analysis sample is equal to 0.5. Results are presented in Fig. 4.

The figure clearly shows that increasing the sample size of a meta-analysis is far more effective in reducing estimator variance and narrowing down confidence intervals. The underlying reason is that deviations of effect sizes from the true underlying value are more and more averaged out when the sample size of the meta-analysis increases. Although these deviations also decrease when the sample size in a primary study increases, they are averaged out to a far lesser extent when the sample size of the meta-analysis remains relatively small. Of course, these results do not imply that the sample size of primary studies does not matter for the outcome of a meta-analysis—it does, especially when sample size is very small. However, the results do show that relatively large meta-analyses with underlying primary studies with a relatively small number of observations are more efficient and produce narrower confidence intervals than relatively small meta-analyses with underlying studies with a relatively small number of observations.

6 Discussion and conclusions

In this study we use Monte-Carlo analysis to investigate the impact of effect size heterogeneity on the results of a meta-analysis. Specifically, we analyse the performance the OLS, the WLS and the mixed effects meta-estimators under three sources of effect size heterogeneity, i.e., heterogeneity in effect size variance (heteroskedasticity) and two types of random effect size variation. For the first type of random effect size variation we replicate the standard assumption in meta-analysis that the random variation holds for all effect sizes in the meta-analysis sample, in which case the mixed effects estimator is the theoretically preferred estimator. For the second type of random variation we abandon this assumption and replicate the more realistic situation in which differences between primary studies, for instance with respect to data type, econometric technique or model specification, cause random variation in the true underlying effect for only part of the effect sizes in a meta-analysis sample. In this case all three meta-estimators use erroneous weights, implying that it is unclear a priori which estimator is preferable. Ultimately, we address the small sample performance of the estimators using the bias, the MSE and the size and power of the statistical tests as performance indicators.

Our results show that increasing heteroskedasticity has a detrimental effect on especially the performance of the OLS estimator. Although the bias is not affected, OLS variance declines considerably compared to its WLS and mixed effects counterparts when heteroskedasticity increases. This pattern changes when we allow for random variation of the true underlying effect across primary studies. Increasing the variance

[14] Although estimator variances go down and the estimators converge in terms of power, this does not mean that the differences between the estimators identified in Sect. 4 disappear. Especially with respect to the MSE the patterns remain unchanged.

of the effect size population increases the variance of all three estimators, but especially that of the WLS estimator. This is due to the fact that WLS has an increasingly erroneous weight structure under these circumstances. Together with the fact that WLS has a downward biased variance estimator under random effect size variation, this leads to a slight increase in size. Clearly, in this case the mixed effects estimator is to be preferred. However, when we induce random variation for only a subset of the effect sizes in the meta-analysis, which may be more realistic, this pattern changes fundamentally. The bias of the WLS estimator remains unaffected, but the bias of the OLS and mixed effects estimators increases when the proportion of effect sizes from misspecified models in the meta-analysis is larger. Admittedly this result is solely due to an interaction effect of omitted variables and erroneously measured effect sizes. Still, since the inclusion of erroneous effect sizes measures in meta-analysis is common, WLS may be preferable in terms of estimator bias to the OLS and mixed effects estimators in many applications in economics. Moreover, the variances of the OLS and mixed effects estimators increase compared to their WLS counterpart, leading to higher MSEs. Furthermore, the size of the statistical tests are at their nominal levels, while the power associated with OLS and mixed effects decreases rapidly for high proportions of misspecification. In reality the source of random effect size variation, if present, is unknown. Because differences between primary studies may induce different types of random variation, it is not unlikely that the variation holds for only a subset of the meta-analysis sample. Our findings show that using the mixed effects estimator in empirical applications of meta-analysis is suboptimal under these circumstances, and that applying WLS is clearly preferable.

Finally, given that meta-estimators have two types of asymptotics, we show that meta-analysis sample size is far more effective in reducing meta-estimator variance and increasing the power of hypothesis testing than primary study sample size. Even for relatively small increases in meta-analysis sample size, the quality of the outcome of a meta-analysis is substantially improved. The crucial factor here is that random effect size deviations from the true underlying effect are averaged out more and more under increasing meta-analysis sample size. Therefore, although the various types of effect size heterogeneity may still have substantial detrimental effects on the small sample performance of meta-estimators, deviations from the true underlying effect average out at sample sizes that are very common in practice.

Acknowledgments This research is supported through the program 'Stimulating the Adoption of Energy-Efficient Technologies', funded by the Netherlands Organization for Scientific Research (NWO) and the Dutch Ministry of Economic Affairs (SenterNovem). We are grateful to two anonymous referees for useful comments on an earlier version of this article. The usual disclaimer applies.

References

Abreu M, De Groot HLF, Florax RJGM (2005) A meta-analysis of beta-convergence: the legendary 2%. J Econ Surv 19:389–420

Bijmolt THA, Pieters RGM (2001) Meta-analysis in marketing when studies contain multiple measurements. Mark Lett 12:157–169
Brockwell SE, Gordon IR (2001) A comparison of statistical methods for meta-analysis. Stat Med 20:825–840
Brons MRA, Nijkamp P, Pels E, Rietveld P (2008) A meta-analysis of the price elasticity of gasoline demand: a SUR approach. Energy Econ 30:2105–2122
De Dominicis L, Florax RJGM, De Groot HLF (2008) Meta-analysis of the relationship between income inequality and economic growth. Scott J Political Econ 55:654–682
Field AP (2001) Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. Psychol Methods 6:161–180
Greene WH (2000) Econometric analysis, 4th edn. Prentice-Hall, Upper Saddle Rivern, New Jersey
Hedges LV (1994) Fixed effects models. In: Cooper H, Hedges LV (eds) The handbook of research synthesis. Russell Sage Foundation, New York
Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic Press, Orlando, Florida
Higgins JPT, Thompson SG (2004) Controlling the risk of spurious findings from meta-regression. Stat Med 23:1663–1682
Hunt M (1997) How science takes stock: the story of meta-analysis. Russell Sage Foundation, New York
Koetse MJ, De Groot HLF, Florax RJGM (2008) Capital-energy substitution and shifts in factor demand: a meta-analysis. Energy Econ 30:2236–2251
Koetse MJ, De Groot HLF, Florax RJGM (2009) A meta-analysis of the investment-uncertainty relationship. South Econ J 76:283–306
Koetse MJ, Florax RJGM, De Groot HLF (2005) Correcting for primary study misspecifications in meta-analysis. Tinbergen Institute Discussion Paper 05-029/3, Tinbergen Institute, Amsterdam
Kuhnert R, Böhning D (2007) A comparison of three different models for estimating relative risk in meta-analysis of 3 clinical trials under unobserved heterogeneity. Stat Med 26:2277–2296
Nijkamp P, Poot J (2004) Meta-analysis of the effect of fiscal policies on long-run growth. Eur J Political Econ 20:91–124
Nijkamp P, Poot J (2005) The last word on the wage curve? A meta-analytic assessment. J Econ Surv 19:421–450
Oswald FL, Johnson JW (1998) On the robustness, bias, and stability of statistics from meta-analysis of correlation coefficients: some initial Monte Carlo findings. J Appl Psychol 83:164–178
Roberts CJ, Stanley TD (2005) Meta-regression analysis: issues of publication bias in economics. Blackwell, Oxford
Sanchez-Meca J, Marin-Martinez F (1997) Homogeneity tests in meta-analysis: a Monte Carlo comparison of statistical power and Type I error. Qual Quant 31:385–399
Sanchez-Meca J, Marin-Martinez F (1998) Weighting by inverse variance or by sample size in meta-analysis: a simulation study. Educ Psychol Meas 58:211–220
Stanley TD (2001) Wheat from chaff: meta-analysis as quantitative literature review. J Econ Perspect 15:131–150
Stanley TD (2008) Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. Oxf Bull Econ Stat 70:103–127
Stanley TD, Jarrell SB (1989) Meta-regression analysis: a quantitative method of literature surveys. J Econ Surv 3:54–67
Sutton AJ, Abrams KR, Sheldon TA, Song F (2000) Methods for meta-analysis in medical research. Wiley, New York
Weichselbaumer D, Winter-Ebmer R (2005) A meta-analysis of the international gender wage gap. J Econ Surv 19:479–511